CS-TR-4736 JULY 2005
LAMP-TR-124
UMIACS-TR-2005-42

# USING WSD TECHNIQUES FOR LEXICAL SELECTION IN STATISTICAL MACHINE TRANSLATION

Clara Cabezas and Philip Resnik

Institute for Advanced Computer Studies
Language and Media Processing Laboratory
Computational Linguistics and Information Processing Laboratory
University of Maryland
College Park, MD 20742-3275
{*clarac,resnik*}*@umiacs.umd.edu*

**Abstract**

In current state of the art statistical MT systems, word choice in the target language is governed implicitly by a combination of "phrase" selection and langage modeling. In contrast, the state of the art in word sense disambiguation takes advantage of a wide array of features, both locally and at the document level. This technical report describes our initial efforts to employ the power of WSD techniques in helping to guide a state of the art statistical MT system toward better word choices.

We briefly discuss the principles underlying our approach as contrasted with another recent attempt to integrate WSD with statistical MT (Carpuat and Wu, 2005) that yielded negative results. We then describe our approach, which leads to a small improvement in translation performance over a state of the art phrase-based statistical MT system. Qualitative analysis of translation output suggests there are still significant opportunities to improve performance further.

This work has been supported in part by ONR MURI Contract FCPO.810548265 and Department of Defense contract RD-02-5700.

# Report Documentation Page

*Form Approved*
*OMB No. 0704-0188*



| | | |
|---|---|---|
| 1. REPORT DATE<br>**JUL 2005** | 2. REPORT TYPE | 3. DATES COVERED<br>**00-07-2005 to 00-07-2005** |

| | |
|---|---|
| 4. TITLE AND SUBTITLE<br>**Using WSD Techniques for Lexical Selection in Statistical Machine Translation** | 5a. CONTRACT NUMBER |
| | 5b. GRANT NUMBER |
| | 5c. PROGRAM ELEMENT NUMBER |
| 6. AUTHOR(S) | 5d. PROJECT NUMBER |
| | 5e. TASK NUMBER |
| | 5f. WORK UNIT NUMBER |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)<br>**Language and Media Processing Laboratory,Institute for Advanced Computer Studies,University of Maryland,College Park,MD,20742-3275** | 8. PERFORMING ORGANIZATION REPORT NUMBER |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) | 10. SPONSOR/MONITOR'S ACRONYM(S) |
| | 11. SPONSOR/MONITOR'S REPORT NUMBER(S) |

12. DISTRIBUTION/AVAILABILITY STATEMENT
**Approved for public release; distribution unlimited**

13. SUPPLEMENTARY NOTES

14. ABSTRACT

15. SUBJECT TERMS

| 16. SECURITY CLASSIFICATION OF: | | | 17. LIMITATION OF ABSTRACT | 18. NUMBER OF PAGES | 19a. NAME OF RESPONSIBLE PERSON |
|---|---|---|---|---|---|
| a. REPORT | b. ABSTRACT | c. THIS PAGE | | **11** | |
| **unclassified** | **unclassified** | **unclassified** | | | |

**Standard Form 298 (Rev. 8-98)**
Prescribed by ANSI Std Z39-18

## 1 Introduction

Statistical machine translation is widely viewed as involving two central problems: the problem of selecting the words in the target language, and the problem obtaining appropriate word order. In the IBM models of the early 1990s (Brown et al., 1990), selection of target-language lexical items was governed by a combination of two main components: a table of lexical translation probabilities $\Pr(f_i|e_j)$ for words $f_i$ and $e_j$, and the language model, determining target-language probability $\Pr(e_1 \ldots e_n)$ typically using an $N$-gram model. Context played a role in lexical disambiguation primarily monolingually, with the language model biasing the decoder toward selecting a word $e_j$ congruent with the previous $N-1$ words in the target-language hypothesis. Bilingual influences on lexical selection did not directly involve any context external to the word being translated, although summing over alternative alignments did allow probabilities $\Pr(f_{i'}|e_j)$ to influence the selection of $e_j$ for $f_{i'}$ elsewhere in the sentence, via an indirect sort of "triggering" effect. Since translation was done strictly on a sentence-by-sentence basis, document-level context played no role at all in influencing word choice.

Since the late 1990s, phrase-based statistical translation models have represented the state of the art in statistical MT (e.g. (Och and Ney, 2004; Koehn et al., 2003; Marcu and Wong, 2002; Kumar et al., 2005)). A phrase table relates contiguous word sequences $\bar{f}$ and $\bar{e}$, capturing both local reordings and local constraints on word choice — for example, Spanish-English phrase correspondences would assign high probability to the relationship between *a escala mundial* and *on a global scale*. In addition to the adjective-noun reordering, the phrase table captures the fact that, in the context of translating this Spanish phrase, *scale* is a more typical English translation choice compared to closely related words such as, say, *magnitude*. This choice is, of course, likely to be reinforced monolingually by the language model, since the probability of *scale* given *on a global...* is high.[1]

Statistical phrases are a positive step in lexical selection for statistical MT, in that they help take better advantage of local context — long known to be an influential factor in determining word meaning in context (Yarowsky, 1993). However, research on the determination of word meaning in context has been converging on the idea that there are actually a whole variety of features that can play a role. In the 2004 SENSEVAL-3 exercise (Mihalcea and Edmonds, 2004), word sense disambiguation systems took advantage not only of string-local features, but also of local part-of-speech information, sentence-level grammatical collocates, and less local or document-level features such as document topic codes, co-occurring words and $N$-grams, and so on. Consistent with Yarowsky and Florian's (2002) observations that no single classifier is a one-size-fits-all solution, the best performing systems took advantage of feature variety and classifier combination approaches. This naturally raises the question of whether the same techniques could have advantages in lexical selection for statistical MT, which bears a very close relationship to monolingual WSD (e.g. see (Dagan and Itai, 1994; Resnik and Yarowsky, 1999; Diab and Resnik, 2002)).

We are undertaking to explore this question, and this technical report presents a picture of where the research currently stands. Our research has been guided thus far by several general considerations.

[1]See Edmonds and Hirst (2002) for a knowledge-based approach to the closely related problem of selecting among near-synonyms in natural language generation.

- First, it is clear from the SENSEVAL exercises that the most successful WSD techniques take advantage of supervised training, and the more data the better.

- Second, our experience in MT suggests caution when attempting to exploit data and knowledge resources external to the bilingual training materials, such as separate WSD-training corpora or previously defined sense inventories. As a case in point, among systems represented in the NIST MT evaluations, the only statistical MT system to exploit hierarchical syntax successfully — the University of Maryland's Hiero system (Chiang, 2005) — learns its synchronous context-free rules directly from the training bitext, rather than trusting a parser trained on an external corpus.

- Third, and related, experience in statistical MT suggests that one should be cautious about turning uncertain decisions into hard constraints. Trusting (or, more to the point, *not* fully trusting) the one-best output of source language parser is one example. Another is Franz Och's use of rule-based translation components, e.g. for dates, numbers, bylines, etc. These are not integrated into his system as hard translation choices, but rather as dynamically generated phrase table entries that can be weighed by the decoder in the context of the entire search.[2]

We believe these three considerations are most likely responsible for the negative (but nonetheless quite interesting) results recently reported by Carpuat and Wu (2005). In their (admittedly first-pass) attempt to integrate state of the art WSD with a Chinese-English MT system, they used a WSD system trained on a relatively small dataset (about 37 training instances per target word), their training dataset and and sense inventory were unrelated to the bilingual training data, and they integrated WSD output via hard constraints (either forcing a choice among WSD-derived candidates at decoding time, or replacing target words in postprocessing).

In the work reported here, we are using target vocabulary items directly as "senses", thus bypassing entirely the question of an externally defined sense inventory. To the extent that word-level alignments are accurate (a non-trivial question, of course), aligned bitext can provide large quantities of material to train from — for example, a sentence pair containing *escala* aligned with *scale* provides a training instance for the former word "tagged" as the latter. As discussed below, we integrate WSD choices as soft decisions by taking advantage of a phrase-based statistical MT system (Koehn, 2004) that optionally permits the specification of confidence-weighted translation alternatives in the source-language input, giving the decoder the choice of whether to use the specified translations or those suggested by its translation model.

In Section 2, we briefly describe our WSD framework, and how it has been adapted for lexical selection in a phrase-based statistical MT system. In Section 3, we describe our preliminary experiments, including quantitative evaluation, qualitative analysis, and thoughts on additional improvements. Section 4 summarizes and concludes.

## 2 Using WSD for Lexical Selection

Beyond the general considerations outlined in the Introduction, our priority in using WSD techniques for lexical selection is a flexible infrastructure that permits an active cycle of experimenta-

[2]Source: Presentation at NIST Machine Translation Workshop, June 20-21, 2005.

tion, data analysis, and algorithm refinement. For that reason, we use the UMD-SST supervised WSD system (Cabezas et al., 2001; Cabezas et al., 2004), which is based on a support vector machine classifier and supports a wide variety of local and less local contextual features.

### 2.1 WSD Infrastructure

The UMD-SST WSD system is described in detail in our SENSEVAL workshop publications (Cabezas et al., 2001; Cabezas et al., 2004). Briefly, the system follows the classic supervised learning paradigm, using a single SVM classifier; each word in the vocabulary is considered an independent classification problem. First, annotated training instances for the ambiguous word are analyzed so that each instance can be represented as a collection of feature-value pairs labeled with the correct category. Then, these data are used for parameter estimation within the supervised learning framework in order to produce a trained classifier. Finally, the trained classifier is given previously unseen test instances and for each instance it yields a confidence score for each of the possible category labels.

Contextual features available in the current system include local collocational features within a window of plus-or-minus 3 words, grammatical collocations within the sentence, and unigrams found within a given extra-sentential context. Features are weighted using *inverse category frequency* (ICF), which is, by analogy with inverse document frequency (IDF), a function of how many distinct categories a feature appears with in training data. Features that occur with most senses of a word have low ICF; those more heavily skewed toward fewer senses have high ICF. In disambiguating a word $w$ with senses $\mathcal{S} = \{s_1, s_2, \ldots, s_{N_w}\}$, we define $\mathrm{ICF}_w(f) = -\log(N_w^f/N_w)$ where $N_w^f$ is the number of distinct elements of $\mathcal{S}$ that ever co-occur with feature $f$ in the training data for word $w$. For example, if a word has five senses, and the feature $L_1$:`the` appears in some training instance for each of the five senses, then $\mathrm{ICF}_w(L_1\text{:}\texttt{the}) = -\log(5/5) = 0$, correctly indicating that this feature is not at all useful for disambiguating among the five senses of this word.

### 2.2 Adapting WSD Techniques for Lexical Selection

Adapting UMD-SST for lexical selection in MT involves a straightforward recasting of aligned target-language words as sense tags. A bilingual corpus is aligned using standard off-the-shelf tools (GIZA++), using English as the target language. The set of "sense tags" for a word is the set of English words with which it is aligned, possibly filtered (see Section 3 for details).

In the current adaptation of the system for lexical selection, local collocation features are defined using all words within a three-word window of the target word, and wide-context features are defined using all words within the current, previous, and following sentence. Grammatical features are not used. Consider the following example as a source of training items.

F. estoy de acuerdo con él en cuanto al <u>papel</u> central que debe conservar en el futuro la comisión como garante del interés general comunitario

E. i agree with him that the commission must continue to play a pivotal <u>role</u> as guardian of the common interests of the community

In this sentence pair, consider the Spanish word *papel*, aligned with *role*. In training a classifier for disambiguating the word *papel*, the sense label would be the English word ROLE, and features from context would include the following:

- Local collocates: {*L3:en, L2:cuanto, L1:al, R1:central, R2:que, R3:debe*}
- Wide-context features: Every token in this sentence and the previous and following sentences. These include *estoy*, *de*, *acuerdo*, etc.

This example illustrates an essentially homonymic distinction: the word *papel* has frequent translations either as *role* or as *paper*. The sentence also contains a nice example of a finer-grained distinction: *central* is here aligned with *pivotal*, though it is also frequently translated as *central*, and sometimes *key*, *decisive*, etc.

### 2.3 Integrating WSD into MT Decoding

The baseline MT system in our experiments was Pharaoh (Koehn, 2004). In addition to being representative of current phrase-based statistical MT approaches, and therefore a proper baseline for comparison, Pharoah makes it possible to investigate the impact of alternative lexical selection decisions while keeping the rest of the translation framework constant. In particular, the Pharaoh decoder allows the option of including, within the source sentence, XML markup indicating translation possibilities for any given span of words in the input. This can be useful for hard rewrites — e.g. forcing European number formats like 3,14159 to be rendered as American 3.14159. More to the point, XML markup can be used to provide soft alternatives, which the decoder will consider along with the alternatives posed by the translation model, the final determination being made by the language model.

As an example, consider the following Spanish input:

> sin embargo , señor presidente también es realmente necesario que en se vaya poco más lejos...

After WSD has applied, the input to the decoder might be:

```
<n english="without|even|no" prob="...">sin</n>
<n english="but|embargo|yet" prob="...">embargo</n>
,
<n english="sir|gentleman|mister" prob="...">se\~nor</n>
<n english="president|chair|speaker" prob="...">presidente</n>
,
<n english="also|too|even" prob="...">tambi\'en</n>
<n english="is|it|be" prob="...">es</n>
<n english="really|indeed|actually" prob="...">realmente</n>
<n english="need|necessary|must" prob="...">necesario</n>
<n english="that|to|than" prob="...">que</n>
<n english="in|on|and" prob="...">en</n>
biarritz
<n english="be|is|been" prob="...">se</n>
<n english="go|goes|going" prob="...">vaya</n>
```

```
un
<n english="little|bit|some" prob="...">poco</n>
<n english="more|over|further" prob="...">m\'as</n>
<n english="far|away|afar" prob="...">lejos</n>
...
```

For the sake of readability, "..." appears here in lieu of probability distributions, and the sentence has been broken across multiple lines.

Crucially, the decoder is free to override preferences expressed in the XML markup, e.g. translating the phrase *sin embargo* as "nevertheless" rather than being forced into something more awkward like "even yet".[3] At the same time, the choice between "Mr. Speaker" and "Mr. President" might be one that is undetermined by sentence-level context, but made clear in the context of the entire document, and thus amenable to a nudge in the right direction from WSD techniques that take advantage of document-level context.

## 3 Preliminary Experimentation

Our preliminary experiments have been conducted using Spanish-English Europarl corpus (Koehn, 2003), randomly sampling 70000 word-aligned sentences for training, 2000 for development, and 2000 for testing. Classifiers are constructed for all Spanish words — not lemmas, and not just content words. The set of possible English "senses" for a word is the set of English words with which it is ever aligned in the training data, filtered by checking a hybrid manual-statistical dictionary.[4]

### 3.1 Quantitative Results

We used BLEU r1n4 (MTeval version 11a) — that is, a single reference translation (r1) and matching up to 4-grams (n4) — to compare the Pharaoh baseline against Pharaoh with WSD-based lexical selection recommendations. The reference translation was simply the English translation for the Spanish test item in the Europarl test set. The decoder output differed by at least one token for 56% of the items in the test set. Including WSD-based lexical selection provides a BLEU score of 0.2382 as compared to the baseline of 0.2356, a 1.1% relative difference.

### 3.2 Qualitative Discussion

Although the improvement in BLEU score is small, and most likely not statistically significant, it is an improvement rather than a decrease in performance (cf. (Carpuat and Wu, 2005)). Moreover, looking at the experimental context, and considering the results qualitatively, there are some reasons to be cautiously optimistic about the possibility of improving the results.

First, BLEU with a single reference is very strict, since it requires an exact match between tokens in the MT output and tokens in the reference translation. The decoder using WSD-based

[3] This is accomplished by running Pharaoh with the `-bypass` flag.

[4] We are grateful to Nizar Habash for providing the manual portion of the dictionary. Statistically derived entries were obtained by computing the log-likelihood ratio for aligned word pairs $\langle e, f \rangle$ in the training data, sorting, and keeping the 100K entries for which the log-likelihood ratio was highest.

lexical selection appears to be making some changes that should be considered improvements, but which are not counted under this strict criterion. For example, consider:

SRC. se sabe por ejemplo que en francia la cifra de ingresos fiscales varia en función de que se tomen las estadísticas de la direción general de contabilidad pública o las de la contabilidad nacional.

REF. it is known that in france , for instance , the figure for tax receipts varies according to whether you use the statistics of the direction générale de la comptabilité publique or those of the comptabilité nationale .

PHA. is in france , for example , the number of tax revenue varies according to take the statistics of the directorate general of the public accounts or of the national accounts .

WSD. for example , we know that the figure in france of income tax varies according to take the statistics of the directorate general of the public accounts or of the national accounts .

In this item, the WSD prediction suggests that *sabe* should be translated as *know*, which presumably helps guide the decoder toward translating *se sabe* as *we know* — this is a perfectly reasonable translation, even though the reference uses *it is known*. This example also illustrates the correct choice of *figure* rather than *number*.

A sampling of other cases where the WSD-enabled lexical selection improves on Pharaoh, but makes reasonable but non-matching choices includes *alleviate the burden* (versus *relieve the burden* in the reference translation), *the duty to remember* (versus *the duty of remembrance*), *reflect seriously about* (versus *reflect seriously on*), and *a complete success* (versus *a triumph*).

Second, WSD-based guidance on lexical choices affects sentence-level translations more globally, not just at the level of individual words. Consider:

SRC. señor presidente , he votado a favor de esta carta en buena parte por la influencia que nuestro colega ingo friedrich y el profesor herzog han ejercido en su contenido .

REF. mr president , i voted in favour of this charter , not least because of the influence which our colleague , ingo friedrich , and professor herzog have exerted on its content .

BAS. i voted for this in a letter to the influence mr ingo friedrich and professor herzog have exercised their content .

NEW. mr president , voted in favour of the charter in large part by the influence mr ingo friedrich and professor herzog have exercised their content .

The baseline decoder chooses to translate *carta* as *letter* (or perhaps even *in a letter*), which leads to a fragment of the translation, *i voted for this in a letter*, that is perfectly fluent but utterly incorrect. In contrast, by translating *carta* correctly as *charter*, the decoder enabled with WSD-based lexical selection not only gets that word correct, but also creates a main verb phrase that more accurately preserves the meaning of the source, *voted in favor of the charter*. Similarly, better translation of function words sometimes has quite a large effect on the meaning. Consider the distinction between *amendments to* and *amendments on*.

SRC. estamos en contra de las enmiendas sobre la masacre de los armenios precisamente por esa misma razón .

REF. we opposed the amendments on the armenian massacre for exactly this reason .

BAS. we are against the amendments to the massacre of armenians by the exactly the same reason .

NEW. we are against the amendments on the massacre of armenians by the exactly the same reason .

Third, the current version of the system is entirely naïve about the grammatical category of the word being translated, except insofar as local collocational features provide stronger evidence for translations in one category versus another. In some cases where WSD-based lexical selection makes incorrect choices, conditioning on grammatical category might provide a better distribution over translations. As an example, the WSD-based preference below for *seguros* as *sure*, rather than *insurance*, leads the decoder to decrease rather than increase the accuracy of the translation.

SRC. en efecto , si cada vez más europeos acuden a los seguros complementarios para ser reembolsados , a para la igualdad de acceso a la asistencia sanitaria .el sector mutualista sigue siendo la mejor garantía para la igualdad de acceso a la asistencia sanitaria .

REF. if more and more europeans turn to supplementary health insurance in order to reimburse health care costs , the mutualist sector will remain the best guarantee for equal access to care .

BAS. if increasingly come to the european supplementary insurance to be reimbursed , the sector mutualista remains the best guarantee for the equal access to health care .

NEW. in fact , if increasingly come to the european complementary sure to be reimbursed , the sector mutualista remains the best guarantee for the equal access to health care .

Biasing the translation in favor of a noun interpretation of *seguros* might well lead the WSD-based selection to the correct conclusion, and consideration of a variety of examples, like those shown above, suggests that introducting a bias based on part-of-speech would not hurt in other cases where WSD is already going in the right direction.

In addition, we suspect that with use of wider context — for example, features from the entire document rather than the three-sentence window — there would be more topical evidence for a more specific meaning like *insurance* rather than a lexical choice like *sure* that is more generic and *a priori* more likely.

Fourth, it is worth noting that the current experiment applied WSD-based lexical guidance across the board, in all cases where a distribution could be obtained. But in many cases, sense distributions are so skewed that it is better to simply use the predominant sense or the sense already favored by the decoder, changing this default only when there is strong evidence in favor of doing so. (This is related to one of the reasons WSD has had very limited success in monolingual information retrieval; see Resnik (forthcoming) for discussion of relevant literature.) Taking this

observation into account suggests using confidence assessment techniques and providing WSD-based lexical selection bias only when one can be confident in the choice. One way to do this would be to add guidance only in cases having a high value for $\Pr(\text{WSD} > \text{PHA}|e, f, \text{context})$ where WSD > PHA indicates that WSD-guided lexical selection is correct when Pharaoh's choice is incorrect.

Fifth and finally, lexical selection in Spanish may be easier than in other languages for a phrase-based MT system accomplishing lexical selection using just its phrase table and language model. For a language like Chinese, where there are likely to be more significant word order and grammatical category divergences with English, a larger arsenal of WSD techniques may turn out to have greater advantages over local context alone. Working with a more heterogeneous corpus than Europarl might have a similar effect.

## 4 Conclusions

In this technical report we have proposed, for the first time, an integration of WSD techniques with statistical phrase-based translation by treating target-language lexical items as "senses". Doing so enables us to take advantage of existing WSD systems by using large aligned bitexts as a source of training data for supervised approaches, and although these data are noisy, all manner of sample selection techniques are therefore available as ways to improve training data quality.

Work still needs to be done in order to obtain real benefits from applying WSD techniques in MT. But our small positive (or at least not negative) result is reassuring, particularly since our baseline system is stronger than the baseline statistical MT system used by Carpuat and Wu's (2005) experiment. We hope to gain from the insights in their careful analysis of negative results, and in the near future we would like to conduct experiments with Chinese in order to obtain a direct comparison of approaches that are and are not mediated by a Chinese word sense inventory.

## Acknowledgments

This work has been supported in part by ONR MURI Contract FCPO.810548265 and Department of Defense contract RD-02-5700. Jessica Stevens contributed to the original implementation of UMD-SST. We are grateful to Franz Och, Philipp Koehn, Thorsten Joachims, and the developers of WEKA (http://www.cs.waikato.ac.nz/~ml/weka/) for publically available software and data used in this research, and to our colleagues in the UMD CLIP Laboratory for their contributions to and support of the machine translation infrastructure.

## References

P. Brown, J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

Clara Cabezas, Philip Resnik, and Jessica Stevens. 2001. Supervised sense tagging using support vector machines. In *Proceedings of the Second International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-2)*, Toulouse, France, July.

Clara Cabezas, Indrajit Bhattacharya, and Philip Resnik. 2004. The university of maryland senseval-3 system descriptions. In *Proceedings of the Third International Workshop on Evaluating Word Sense Disambiguation Systems (SENSEVAL-3)*, July.

Marine Carpuat and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005)*, June.

David Chiang. 2005. A hierarchical phrase-based model for statistical machine translation. In *Proceedings of ACL 2005*, pages 263–270.

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *40th Anniversary Meeting of the Association for Computational Linguistics (ACL-02)*, Philadelphia, July.

Philip Edmonds and Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics*, 28(2):105–144.

Philipp Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *HLT/NAACL-03*, pages 127–133.

Philipp Koehn. 2003. Europarl: A multilingual corpus for evaluation of machine translation. Unpublished draft, http://people.csail.mit.edu/koehn/publications/europarl/.

Philipp Koehn. 2004. Pharaoh, a beam search decoder for phrase-based statistical machine translation models: User manual and description for version 1.2, August. http://www.isi.edu/licensed-sw/pharaoh/manual-v1.2.ps.

Shankar Kumar, Yonggang Deng, and William Byrne. 2005. A weighted finite state transducer translation template model for statistic al machine translation. *JNLE*. To appear.

Daniel Marcu and William Wong. 2002. A phrase-based, joint probability model for statistical machine translation. In *EMNLP-2002*, pages 133–139.

Rada Mihalcea and Phil Edmonds, editors. 2004. *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Association for Computational Linguistics, Barcelona, Spain, July.

Franz Josef Och and Hermann Ney. 2004. The alignment template approach to statistical machine translation. *Computational Linguistics*, 30:417–449.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: New evaluation methods for word sense disambiguation. *Natural Language Engineering*, 5(2):113–133.

Philip Resnik. forthcoming. Word sense disambiguation in nlp applications. In Eneko Agirre and Philip Edmonds, editors, *Word Sense Disambiguation: Algorithms and Applications*. Kluwer.

David Yarowsky and Radu Florian. 2002. Evaluating sense disambiguation across diverse parameter spaces. *Natural Language Engineering*, 8(4):293–310.

David Yarowsky. 1993. One sense per collocation. ARPA Workshop on Human Language Technology, March. Princeton.